Density Estimation with Gaussian Mixture Models

In this chapter will introduce important concepts of the *expectation maximisation (EM) algorithm* and a latent variable perspective of *density estimation with mixture models*.

Density estimation represents data compactly using a density from a parametric family, e.g. a Gaussian or Beta distribution, by finding the mean and variance of a dataset. Ways of finding the mean and variance are using maximum likelihood or maximum a posteriori estimation.

In practice, the Gaussian may have limited modeling capabilities, the result may be poor. Therefore, a more expressive family of distributions introducing here is the *mixture models*.

Mixture models can be used to describe a distribution $p(x)$ by a convex combination of K simple distributions

$$\begin{aligned} p(x) &= \sum_{k=1}^{K} \pi_k p_k(x) \\ 0 \leq \pi_k \leq 1, &\sum_{k=1}^{K} \pi_k = 1 \end{aligned} \tag{1}$$

where the components p_k are members of a family of basic distributions and π_k are mixture weights.

Mixture models are more expressive because they allow for multimodal data representations, they can describe datasets with multiple “clusters”.

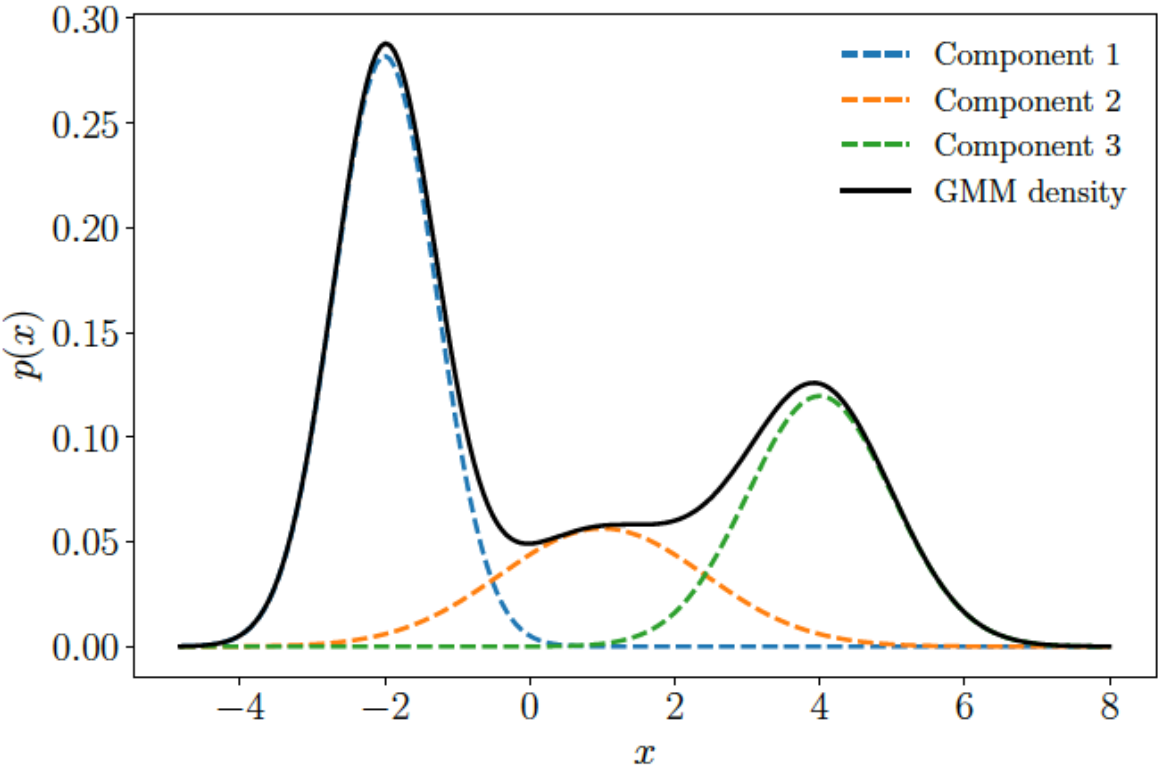
Gaussian Mixture Model

A *Gaussian Mixture Model* (GMM) is a density model which combines a finite number of K Gaussian distributions $\mathcal{N}(x|\mu_k, \Sigma_k)$ so that

$$
\begin{aligned}
p(x|\theta) &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \\
0 \leq \pi_k \leq 1,& \sum_{k=1}^{K} \pi_k = 1
\end{aligned}
\tag{2}
$$

where defined $\theta := \{\mu_k, \Sigma_k, \pi_k : k = 1, \ldots, K\}$ as the collection of all parameters of the model, π_k is the mixture weights.

This combination of Gaussian distribution is more flexibility for modeling complex densities than a simple Gaussian distribution for $K = 1$.

> An example of weighted components and the mixture density, which given as
>
> $$p(x|\theta) = 0.5\mathcal{N}(x| -2, \frac{1}{2}) + 0.2\mathcal{N}(x|1, 2) + 0.3\mathcal{N}(x|4, 1)$$
>
> 
>

Parameter Learning via Maximum Likelihood

Assume a dataset $\mathcal{X} = \{x_1, \ldots, x_N\}$, where x_n are drawn i.i.d. from an unknown distribution $p(x)$.

The objective is to find a good approximation/representation of the unknown distribution $p(x)$ by a GMM with K mixture components.

The parameters of the GMM are K means μ_k, the covariances Σ_k, and mixture weights π_k, summarising to $\theta := \{\pi_k, \mu_k, \Sigma_k : k = 1, \ldots, k\}$.

How to obtain a maximum likelihood estimate θ_{ML}

Write down the likelihood, the predictive distribution of the training data given the parameters. Using i.i.d. assumption,

$$p(\mathcal{X}|\theta) = \prod_{n=1}^{N} p(x_n|\theta),\ p(x_n|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

then obtain the log-likelihood as

$$\log p(\mathcal{X}|\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) = \underbrace{\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}_{:=\mathcal{L}}$$

The normal way to find parameters θ^*_{ML} is to compute the gradient $d\mathcal{L}/d\theta$ of the log-likelihood to θ, set to 0, and solve for θ.

But there's no closed-form solution here since we cannot move the $\log$ into the sum over k.

The EM algorithm for GMMs

An iterative scheme to find good model parameters θ_{ML}, and the key idea is to update one model parameter at a time while keeping the others fixed.

To optimise the log-likelihood with respect to the GMM parameters:

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = 0^T \iff \sum_{n=1}^{N} \frac{\partial \log p(x_n|\theta)}{\partial \mu_k} = 0^T$$
$$\frac{\partial \mathcal{L}}{\partial \Sigma_k} = 0 \iff \sum_{n=1}^{N} \frac{\partial \log p(x_n|\theta)}{\partial \Sigma_k} = 0$$
$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \iff \sum_{n=1}^{N} \frac{\partial \log p(x_n|\theta)}{\partial \pi_k} = 0$$

For all three necessary conditions, by applying the chain rule,

$$\frac{\partial \log p(x_n|\theta)}{\partial \theta} = \frac{1}{p(x_n|\theta)} \frac{\partial p(x_n|\theta)}{\partial \theta}$$

where $\theta = \{\mu_k, \Sigma_k, \pi_k, k = 1, \ldots, K\}$ are the model parameters and

$$\frac{1}{p(x_n|\theta)} = \frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

Responsibilities

Define the quantity

$$r_{nk} := \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

as the *responsibility* of the k-th mixture component for the n-th data point.

The responsibility r_{nk} of the k-th mixture component for data point x_n is proportional to the likelihood

$$p(x_n|\pi_k, \mu_k, \Sigma_k) = \pi_k \mathcal{N}(x_n|\mu_j, \Sigma_j)$$

of the mixture component given the data point.

> Therefore, mixture components have a high responsibility for a data point when the data point could be a plausible sample from that mixture

component.

> 💡 $\mathbf{r}_n := [r_{n1}, \ldots, r_{nK}] \in \mathbb{R}^K$ is a (normalised) probability vector, i.e., $\sum_k r_{nk} = 1$ with $r_{nk} \geq 0$.

Therefore, the responsibility r_{nk} represents the probability that x_n has been generated by the k-th mixture component.

Updating the Means, Covariances, Mixture Weights

Theorem 11.1 Update GMM Means. The update of the mean parameters μ_k of the GMM is

$$\mu_k^{new} = \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}} = \frac{1}{N_k} \sum_{n=1}^N r_{nk} x_n$$

Define that

$$N_k := \sum_{n=1}^N r_{nk}$$

$$L(\theta) = \sum_n \log p(x_n|\theta) = \sum_{n=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \quad \text{Log likelihood}$$

$$\frac{\partial L(\theta)}{\partial \mu_k} = \sum_{n=1}^N \frac{1}{p(x_n|\theta)} \frac{\partial}{\partial \mu_k} \sum_{j=1}^K \pi_j \mathcal{N}(x_n; \mu_j, \Sigma_j) \quad \text{gradient of log likelihood wrt mean}$$

$$= \sum_{n=1}^N \frac{1}{p(x_n|\theta)} \pi_k \frac{\partial}{\partial \mu_k} \mathcal{N}(x_n; \mu_k, \Sigma_k) \quad \text{Ignore terms that do not depend on k}$$

$$= \sum_{n=1}^N \frac{1}{p(x_n|\theta)} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)(x_n - \mu_k)^\top \Sigma_k^{-1} \quad \text{Gradient of Gaussian wrt mean}$$

$$= \sum_{n=1}^N r_{nk} (x_n - \mu_k)^\top \Sigma_k^{-1} \quad \text{Rewrite using responsibility}$$

$$\frac{\partial L(\theta)}{\partial \mu_k} = \mathbf{0}^\top \rightarrow \mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \quad \text{Set gradient to zero}$$

Theorem 11.2 Update GMM Covariances. The update of the covariance parameters Σ_k of the GMM is

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk}(x_n - \mu_k)(x_n - \mu_k)^T$$

$$L(\theta) = \sum_n \log p(x_n|\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)$$ Log likelihood

$$\frac{\partial L(\theta)}{\partial \Sigma_k} = \sum_{n=1}^{N} \frac{1}{p(x_n|\theta)} \frac{\partial}{\partial \Sigma_k} \sum_{j=1}^{K} \pi_j \mathcal{N}(x_n; \mu_j, \Sigma_j)$$ gradient of log likelihood wrt cov

$$= \sum_{n=1}^{N} \frac{1}{p(x_n|\theta)} \pi_k \frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n; \mu_k, \Sigma_k)$$ Ignore terms that do not depend on k

$$\frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n; \mu_k, \Sigma_k) = \frac{\partial}{\partial \Sigma_k} \left[(2\pi)^{-\frac{D}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k) \right) \right]$$ Gradient of Gaussian wrt cov

$$= (2\pi)^{-\frac{D}{2}} \frac{-1}{2} |\Sigma_k|^{-\frac{1}{2}} \Sigma_k^{-1} \exp\left(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k) \right)$$

$$+ (2\pi)^{-\frac{D}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k) \right) \frac{1}{2} \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}$$

$$= -\frac{1}{2} \mathcal{N}(x_n; \mu_k, \Sigma_k)[\Sigma_k^{-1} - \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}]$$

gradient of log likelihood wrt cov

$$\frac{\partial L(\theta)}{\partial \Sigma_k} = \sum_{n=1}^{N} \frac{1}{p(x_n|\theta)} \pi_k \frac{-1}{2} \mathcal{N}(x_n; \mu_k, \Sigma_k)[\Sigma_k^{-1} - \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}]$$

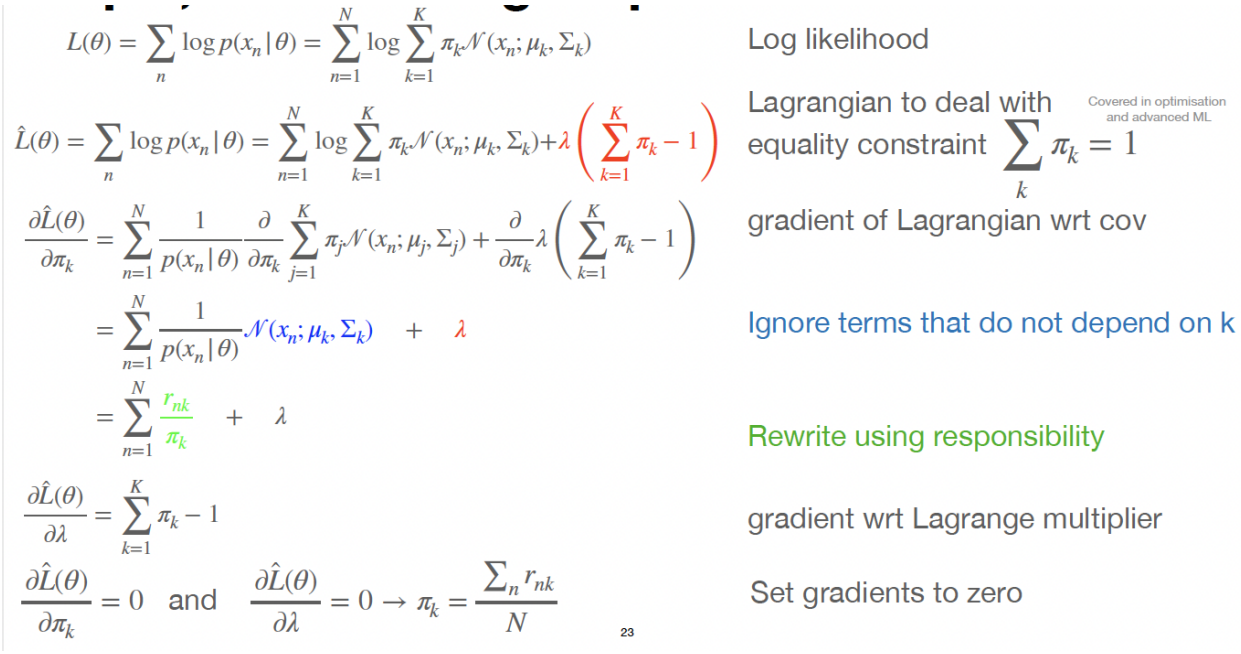
$$= -\frac{1}{2} \sum_{n=1}^{N} r_{nk}[\Sigma_k^{-1} - \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}]$$ Rewrite using responsibility

$$\frac{\partial L(\theta)}{\partial \Sigma_k} = \mathbf{0} \rightarrow \Sigma_k = \frac{\sum_n r_{nk}(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_n r_{nk}}$$ Set gradient to zero

New covariance = weighted covariance of data where weights = responsibilities

Theorem 11.3 Update GMM Mixture Weights. The update of the mixture weights parameters μ_k of the GMM is

$$\pi_k^{new} = \frac{N_k}{N}, \; k = 1, \ldots, K$$



EM Algorithm

The updates above don't have a closed-form solution because the responsibilities r_{nk} depend on those parameters in a complex way.

The Expectation Maximisation algorithm is a general iterative scheme for learning parameters (maximum likelihood or MAP) in mixture models and, more generally, latent-variable models.

Every step in EM algorithm increases the log-likelihood function. For convergence, check the log-likelihood or the parameters directly.

For Gaussian mixture model,

- initialise values for μ_k, Σ_k, π_k
- *E-step*: evaluate the responsibilities r_{nk} for every data point x_n using current parameters

$$r_{nk} := \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

- *M-step*: use the updated responsibilities to reestimate the parameters

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$

$$\pi_k^{new} = \frac{N_k}{N}$$

Latent-Variable Perspective

The goal is to maximum the likelihood, the $\arg\max_\theta p(X|\theta)$, which is equivalent to $\arg\max_\theta \log p(X|\theta)$.

With latent variables $\{z_n\}_{n=1}^N$ representing component assignment

$$L(\theta) = \log p(X \mid \theta) = \log \int p(X, z \mid \theta)\, \mathrm{d}z = \log \int q(z) \frac{p(X, z \mid \theta)}{q(z)}\, \mathrm{d}z \geq \int q(z) \log \frac{p(X, z \mid \theta)}{q(z)}\, \mathrm{d}z := \mathcal{F}(q(z), \theta)$$

An arbitrary $q(z)$ — Jensen's inequality — Lower bound on $L(\theta)$

Instead of maximising $L(\theta)$ directly, we will maximise the lower bound $\mathcal{F}(q(z), \theta)$, alternating between $q(z)$ and θ while keeping the other fixed.

Tips

1. If K takes a greater value, the likelihood becomes greater after convergence.
2. Assume there are N data points, the maximum likelihood will be achieved if $K = N$.
3. GMM has a higher computational complexity than K-means.